Geophysical Research Abstracts
Vol. 17, EGU2015-8273, 2015
EGU General Assembly 2015

NCI's High Performance Computing (HPC) and High Performance Data (HPD) Computing Platform for Environmental and Earth System Data Science

Ben Evans (1), Chris Allen (1), Joseph Antony (1), Irina Bastrakova (2), Kashif Gohar (1), David Porter (1), Tim Pugh (3), Fabiana Santana (1), Jon Smillie (1), Claire Trenham (1), Jingbo Wang (1), and Lesley Wyborn (1)

(1) NCI, Australian National University, Canberra, Australia (Ben.Evans@anu.edu.au), (2) Geoscience Australia, Canberra, Australia, (3) Australian Bureau of Meteorology, Melbourne, Australia

The National Computational Infrastructure (NCI) has established a powerful and flexible in-situ petascale computational environment to enable both high performance computing and Data-intensive Science across a wide spectrum of national environmental and earth science data collections – in particular climate, observational data and geoscientific assets. This paper examines 1) the computational environments that supports the modelling and data processing pipelines, 2) the analysis environments and methods to support data analysis, and 3) the progress so far to harmonise the underlying data collections for future interdisciplinary research across these large volume data collections.

NCI has established 10+ PBytes of major national and international data collections from both the government and research sectors based on six themes: 1) weather, climate, and earth system science model simulations, 2) marine and earth observations, 3) geosciences, 4) terrestrial ecosystems, 5) water and hydrology, and 6) astronomy, social and biosciences. Collectively they span the lithosphere, crust, biosphere, hydrosphere, troposphere, and stratosphere. The data is largely sourced from NCI's partners (which include the custodians of many of the major Australian national-scale scientific collections), leading research communities, and collaborating overseas organisations.

New infrastructures created at NCI mean the data collections are now accessible within an integrated High Performance Computing and Data (HPC-HPD) environment - a 1.2 PFlop supercomputer (Raijin), a HPC class 3000 core OpenStack cloud system and several highly connected large-scale high-bandwidth Lustre filesystems. The hardware was designed at inception to ensure that it would allow the layered software environment to flexibly accommodate the advancement of future data science.

New approaches to software technology and data models have also had to be developed to enable access to these large and exponentially increasing data volumes at NCI. Traditional HPC and data environments are still made available in a way that flexibly provides the tools, services and supporting software systems on these new petascale infrastructures. But to enable the research to take place at this scale, the data, metadata and software now need to evolve together - creating a new integrated high performance infrastructure.

The new infrastructure at NCI currently supports a catalogue of integrated, reusable software and workflows from earth system and ecosystem modelling, weather research, satellite and other observed data processing and analysis. One of the challenges for NCI has been to support existing techniques and methods, while carefully preparing the underlying infrastructure for the transition needed for the next class of Data-intensive Science. In doing so, a flexible range of techniques and software can be made available for application across the corpus of data collections available, and to provide a new infrastructure for future interdisciplinary research.